

Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee



Lecture – 22
Multiple Linear Regression - Part I

Welcome, to the course Business Analytics and Data Mining Modeling Using R. So, we are about to start our next module that is supervised learning methods and we are going to start our first topic that is first technique that is multiple linear regression. So, let us start. So, multiple linear regression is one of the most popular model that is used for statistical modeling as well as data mining modeling. In most of the text books and other stat related courses this is one of the first model that is generally discussed and covered.

(Refer Slide Time: 01:06)

MULTIPLE LINEAR REGRESSION

- Most popular model
- Idea is to fit a linear relationship between
 - A quantitative outcome variable (Y) and
 - A set of p predictors $\{X_1, X_2, X_3, \dots, X_p\}$
- Assumption: relationship as expressed in the following model equation holds true for the target population
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$
Where β_0, \dots, β_p are coefficients and ϵ is the noise or unexplained part

IT RoorkeeNPTEL ONLINE
CERTIFICATION COURSE2

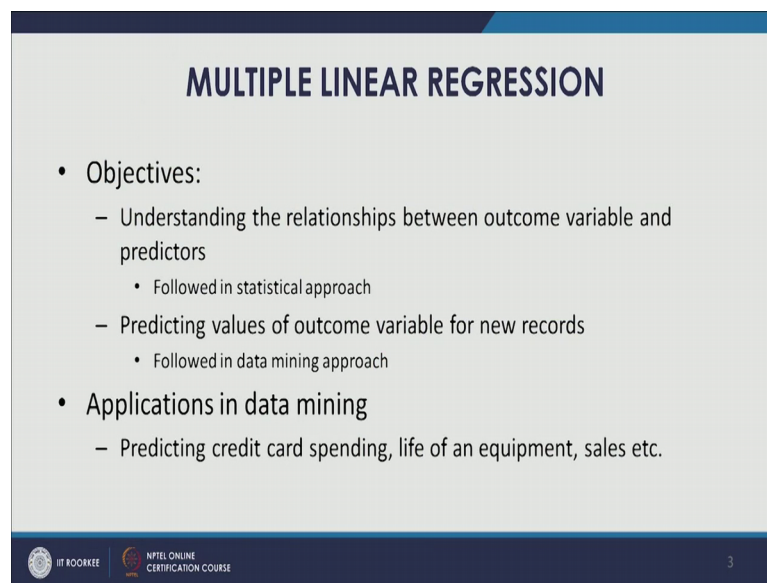
Now, multiple linear regression model the main idea is to fit a linear relationship between a quantitative outcome variable, that is, Y and a set of p predictors for example, X_1, X_2, X_3 and X_p . So, one difference between the statistical techniques and data mining techniques that we have described before as well is that generally in a statistical technique we have assumed relationship between output variable and the set of predictors.

So, for example, in this case multiple linear regression you would see that the first assumption that we are talking about in the slide is that the relationship as expressed in the following model equation that is the linear model equation holds true for the target population so; that means, this relationship linear relationship is actually assumed. So, therefore, when we talk about statistical modeling generally we make certain assumptions about the data, certain assumption about the structure of the data, relationship between variables and we also hypothesize few relationships and then also test them through techniques, through different statistical techniques.

So, the same thing is applicable to multiple linear regression as well and the first assumption is that this as expressed in this particular equation the linear relationship is assumed here. Now, the beta zeros, beta 1, beta 2 and beta p till beta p these are all regression coefficients and then we have this epsilon term that is the noise or unexplained part. So, generally for the exploratory models the predictors information that is used to explain the variability in the outcome variable that is Y and the noise is actually the unexplained part, something that we are not able to explain right with the help of the predictors information that goes into the noise. Now, there could be 2 objectives while where we could actually use them this particular technique multiple linear regression. So, one objective could be understanding the relationships between outcome variable and predictors. This is the typical objective that is followed in a statistical approach. The second objective which is generally followed in a data mining approach is predicting value of outcome variables for new records.

So, as we will discuss in this lecture depending on the objective our approach modeling model building might remain same to a large extent, but the results interpretation the model evaluation that would actually change. So, that is very closely tied to the objective. So, the first objective that we just talked about understanding the relationships this particular objective is mainly explanatory in nature. The second objective the predicting value that is the you know predictive in nature and the predictive modeling would be required.

(Refer Slide Time: 04:42)



MULTIPLE LINEAR REGRESSION

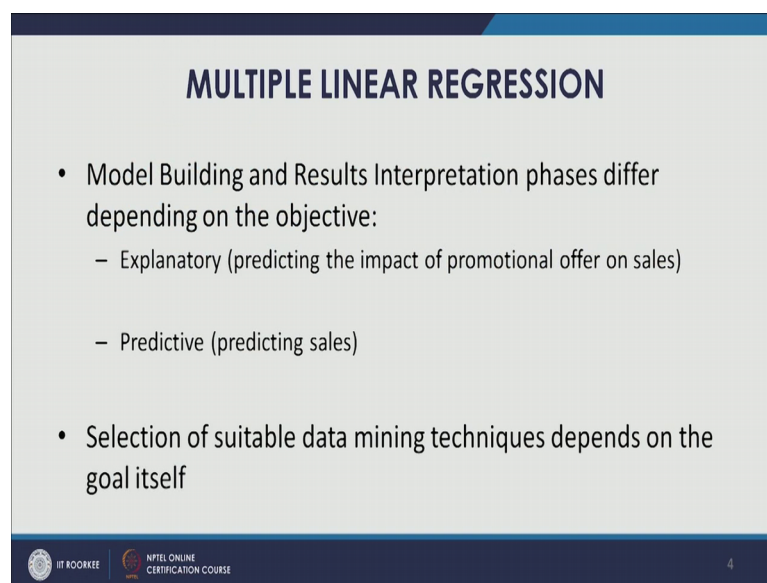
- Objectives:
 - Understanding the relationships between outcome variable and predictors
 - Followed in statistical approach
 - Predicting values of outcome variable for new records
 - Followed in data mining approach
- Applications in data mining
 - Predicting credit card spending, life of an equipment, sales etc.

III ROORKEE NPTEL ONLINE CERTIFICATION COURSE 3

Now, applications in data mining, so, multiple linear regression has many applications in data mining situations. For example, if you predicting credit card spending, predicting life of equipment predicting, sales. So, many such examples that we have been talking about in this particular course. So, multiple linear regression as a technique can be applied in many such situations that we have been talking about. So, mainly multiple linear regression is used to handle the prediction tasks and that is also very well reflected in the previous slide when we said that the outcome variable is quantitative in nature. So, when we said that the model is used the main idea is to fit a linear relationship between this quantitative outcome variable. So, therefore, in a data mining situation essentially when we use multiple linear regression we would be predicting the value of this particular outcome variable therefore, it is the predict prediction task where this particular technique is used.

Now, the goal that we talked about as we talked about that the selection of the model is particularly tied with the goal and it is the model building process might remain same to the large extent, but the results interpretation. So, that will differ depending on the goal. So, which is explained in this particular slide that, for example, predicting the impact of promotional offer on sales.

(Refer Slide Time: 06:25)



MULTIPLE LINEAR REGRESSION

- Model Building and Results Interpretation phases differ depending on the objective:
 - Explanatory (predicting the impact of promotional offer on sales)
 - Predictive (predicting sales)
- Selection of suitable data mining techniques depends on the goal itself

IIIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 4

This would be even though we are trying to predict this impact this in a sense is explanatory task and therefore, the interpretation results interpretation and model building exercise would slightly differ from if we have a goal of just predicting sales. So, that would be predictive goal, predictive task and predictive waddling would be required. So, these 2 task these 2 examples that we just discussed. Predicting the impact of promotional offer on sales which is explanatory and then predicting sales which are predictive. So, the model welding exercise and the later on the model evaluation and result interpretation that would differ from in the between these 2 cases.

Now, as we discussed that selection of a suitable data mining technique it will also depend on the goal itself, whether the goal is explanatory or predictive. To discuss little bit more about, because the multiple linear regression modeling because the most of the statistical technique that we are going to cover in this course. So, they are going to be used in a predictive analytics setting in a data mining modeling. So, there, but they are also used in the statistical setting as well. So, therefore, we would like to again differentiate these 2 environments. So, one is explanatory modelling then other one is predictive modeling. So, let us go through some of the let us understand some of the differences that is there in explanatory modeling and predictive modelling.

(Refer Slide Time: 08:14)

MULTIPLE LINEAR REGRESSION	
Explanatory Modeling	Predictive Modeling
<ul style="list-style-type: none">• Fits the data closely• Full sample is used to estimate best-fit model• Performance metrics measure how close model fits the data	<ul style="list-style-type: none">• Predicts new records accurately• Sample is partitioned into training, validation, and test sets and training partition is used to estimate the model• Performance metrics measure how well model predicts new observations

So, when we do explanatory modeling it is about, we want to find out a model that fits the data closely. So, that is our objective when we talk about predictive modeling we want to find out the model, the best model that predicts new records accurately. Similarly, when we talk about explanatory modeling the full sample that we have that is used to estimate the best fit model, the model that is best fitting the data, the full sample is used there. When we talk about predictive modeling the sample is as we have been talking about in the previous lecture, sample is partitioned into training validation in test set and it is the test it is the training partition that is used to estimate the model. There are other differences for example, performance matrix.

So, the performance matrix in explanatory modeling they measure how close model fits the data. So, the model that we wanted we want wanted a model that fits the data closely and we also require performance matrix which measure the same thing which measure how close model fits the data when we talk about predictive modeling then the performance matrix that we require or performance matrix that we use they should measure how well model predicts new observations right. So, many such matrix we have talked about in our previous lectures on performance matrix.

So, you can see a clear difference between these 2 modeling approaches explanatory modelling and predictive modeling. Let us move forward.

(Refer Slide Time: 10:10)

The slide is titled "MULTIPLE LINEAR REGRESSION" in bold, dark blue text at the top center. Below the title, there are two columns of text. The left column is headed "Explanatory Modeling" and the right column is headed "Predictive Modeling". Each column contains a bulleted list of points. At the bottom of the slide, there is a dark blue footer bar containing the IIT ROORKEE logo, the text "IIT ROORKEE", the NFTEL ONLINE CERTIFICATION COURSE logo, and the text "NFTEL ONLINE CERTIFICATION COURSE". A small number "6" is visible in the bottom right corner of the slide.

Explanatory Modeling	Predictive Modeling
<ul style="list-style-type: none">• Model might not have best predictive accuracy• Statistical techniques with assumed or hypothesized relationships and scarce data (primary data)	<ul style="list-style-type: none">• Model might not be best-fit of data• Machine learning techniques with no assumed structure and large datasets (secondary data)

IIT ROORKEE | NFTEL ONLINE CERTIFICATION COURSE

6

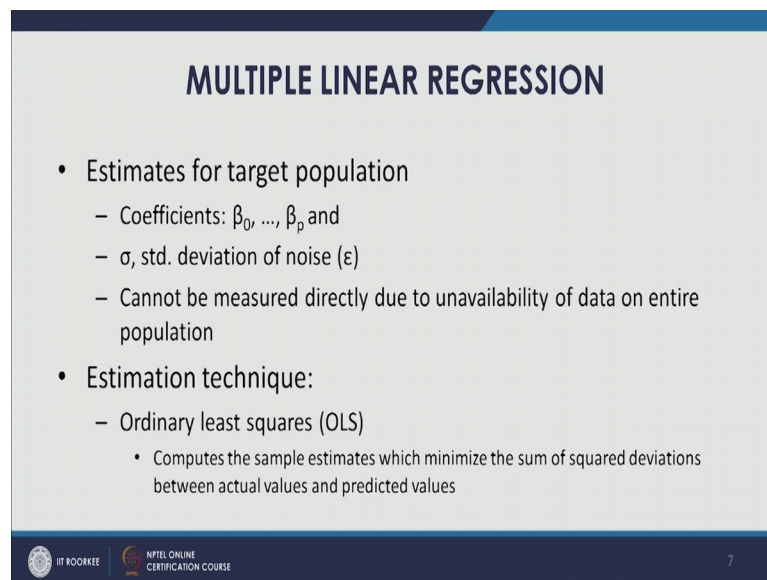
So, there are few more things for example, model that we might select the final model that we might have after this model building process and everything other phases. The model might not have best predictive accuracy because the purpose was as we talked about in the previous slide we want a model that fits the data closely. So, the model that we might get final model might not have best predictive accuracy. Now, we look at the predictive modeling scenario the model that we might finally select out of this exercise in predictive modeling model might not be best fit of data. So, clearly the difference is very clearly we can understand.

Now, to further I understand the explanatory modeling, generally statistical techniques with assumed or hypothesize relationships and then a scarce data which is generally the primary data we are operating in that part. So, we generally use statistical techniques and the assumed or hypothesized re relationships. So, we always depending for a particular problem we always formulate our hypothesis and we test that hypothesis using particular statistical technique by building a model. So, that is the world where we operate and the data is always the sample is the view that is data collection is always difficulty we deal with the scarce data and generally primary data is collected.

Now, we look at the predictive modeling. Typically, we are operating in the machine learning we generally use machine learning techniques and these machine learning techniques they assume you know there is no assumed structure. So, we do not force any

structure on data when we use machine learning techniques and we are generally dealing with large data sets. So, these are generally dealing with secondary data and this data is then because this being large data set we can do other things for example, partitioning which can minimize some of the problems that we face during statistical modeling.

(Refer Slide Time: 12:46)



MULTIPLE LINEAR REGRESSION

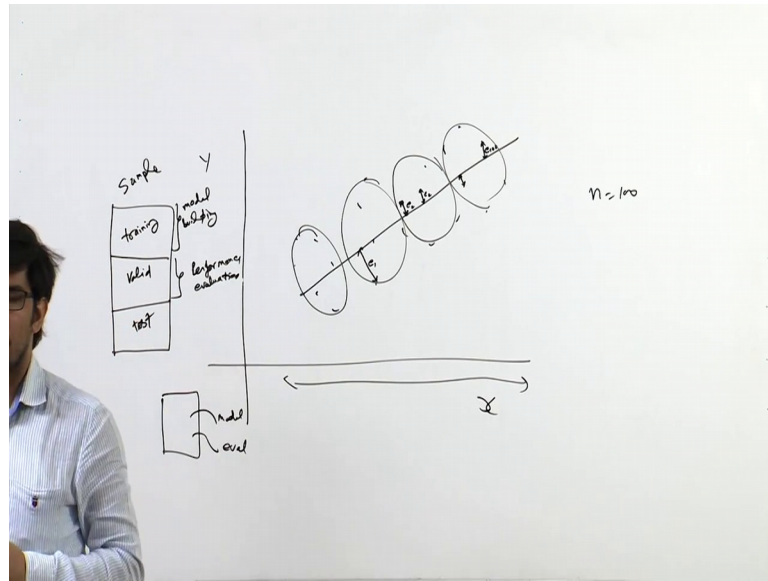
- Estimates for target population
 - Coefficients: β_0, \dots, β_p and
 - σ , std. deviation of noise (ϵ)
 - Cannot be measured directly due to unavailability of data on entire population
- Estimation technique:
 - Ordinary least squares (OLS)
 - Computes the sample estimates which minimize the sum of squared deviations between actual values and predicted values

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 7

Now, the regression equation that we talked about, so we also talked about the regression coefficient beta 0 to beta p; if there are p predictors and then another estimate that we need to compute is this sigma, that is, a standard deviation of noise; noise be denoted using epsilon. So, these estimates we need to compute to find out we need to have about the target population to understand the relationships right first. Now, we these estimates cannot be measured directly, because we do not have the data available on the entire population and that is why we take a sample and it is the sample on which we apply our estimation techniques and we compute these estimates beta 0 to beta p and then sigma that is standard deviation of noise.

So, typically there are many techniques that could be used to estimate these coefficients beta 0, beta 1 to beta p and sigma, but typically ordinary least squares the OLS is the technique that is used to compute these estimates from a sample. So, OLS will compute the sample estimates which minimize the sum of the squared deviations between actual values and predicted values.

(Refer Slide Time: 14:16)



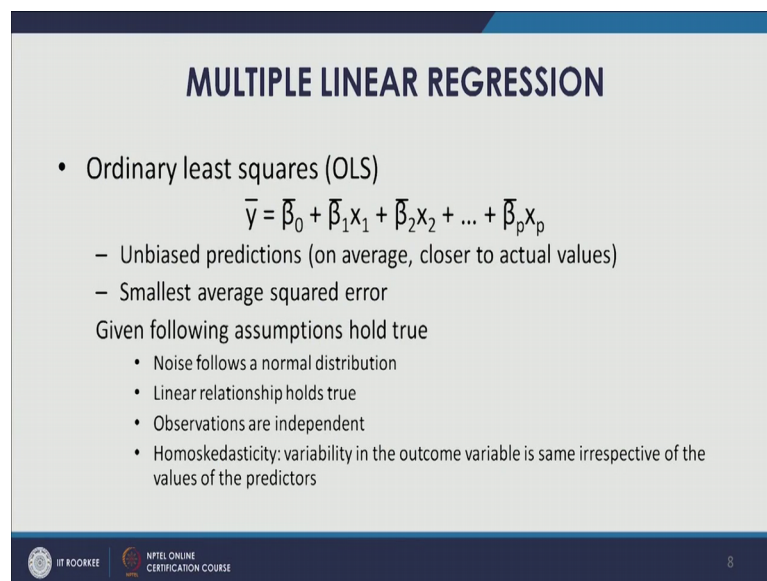
So, let us understand this particular thing by plotting a model. So, let us have we have a this kind of data and the regression equation that we just saw in slides and. So, we were able to find a particular line that would fit this particular data as closely as possible. So, probably this would be the line. So, this is the line that we after applying regression on data set this particular data set these observation we got this line which is closely fitting the data.

Now, how this line has been estimated and this line is also being represented by these estimates β_0 , coefficient and σ . Now, how OLS works at that is very well defined in the slides that we tried try to minimize the sum of squared deviation between actual value and predicted value. So, therefore, all this line will actually have the predicted values. So, actual values we can see on screen. So, all these actual values and these are the actual values and the corresponding predicted value is going to be represented somewhere in the line. So, if we are able to connect the corresponding points the predicted value points on this particular line and the actual points that are their actual observations that are there in this particular graph and we are able to connect, so, this would be the deviations right. So, these would be the errors for individual observations or individual cases.

Now, OLS will always try to minimize the sum of squared deviations. So, let us say you have n observation, let us say we have 100 observations. So, for each observation the

squared value of these errors the actual value minus predicted value has we talked about in our previous lecture performance matrix. So, these squared, these errors actual value minus predicted value this square value of these deviations and then the summation of this would actually minimized by the OLS and that is how this line would be computed. So, the line that we get is computed the estimates that we get the beta 0, beta 1 to beta p and sigma that we get is by following this process we try to minimize these numbers, sum of squared deviations between actual values and predicted values.

(Refer Slide Time: 17:48)





MULTIPLE LINEAR REGRESSION

- Ordinary least squares (OLS)

$$\bar{y} = \bar{\beta}_0 + \bar{\beta}_1 x_1 + \bar{\beta}_2 x_2 + \dots + \bar{\beta}_p x_p$$
 - Unbiased predictions (on average, closer to actual values)
 - Smallest average squared error

Given following assumptions hold true

- Noise follows a normal distribution
- Linear relationship holds true
- Observations are independent
- Homoskedasticity: variability in the outcome variable is same irrespective of the values of the predictors

 IIT ROORKEE
  NPTEL ONLINE CERTIFICATION COURSE
 8

Now, if we want to compute the predicted value for different observations this is how we can do it. You can see on a screen that ordinary least square on this particular slide. That given these values x_1, x_2 given these values on predictors x_1, x_2 and x_p and since we have you know estimated these betas right these beta values. So, because these are sample estimates. So, OLS is applied on the sample, we get the sample estimates. So, these are being represented by $\bar{\beta}_0, \bar{\beta}_1$. So, these being sample estimates. So, we have these numbers and we have the information on the predictors the values on predictor. So, these are going to be used as per this equation and will get the \bar{y} that is the predicted value and any difference between actual value that is y and this \bar{y} that is predicted value that will become the error for that particular record or the deviation.

So, following this method OLS after applying OLS and computing these estimates this β_0 and the standard deviation, we will get we can get unbiased predictions. So, that is for all the observations the predicted values that we get on an average we can assume that these values are going to be closer to actual values, because the idea was to minimize these particular deviation, the idea was to minimize this error and we will also get a smallest average squared error because that is the method that we have followed, but certain assumptions should hold true. So, first assumption is about noise follows a normal distribution.

So, the noise term that we had right in the regression equation it should follow a normal distribution or equivalently we can also say that the outcome variable should be following normal distribution. So, this is the first assumption. Now, we require this assumption mainly in statistical modeling within a statistical modeling when we build our model we build it on the same sample and the reliability of the estimates are also assessed on the same sample therefore, there is therefore, the estimates might not estimates might lack reliability, because in statistical modeling we are always looking for a model which best fits the data. So, that might lead to over fitting. So, therefore, the estimates might not be reliable. So, therefore, we need to draw confidence intervals to have a range and then we should be able to claim that those values would fall within those within those ranges.

So, therefore, for us to be able to compute those confidence intervals, those regions we need to have this first assumption have has to be held true that is noise should follow normal distribution only then we would be able to arrive derive those confidence intervals and therefore, we would be able to claim that the those particular estimates whether it is about the mean of population or anything else they would fall within that particular range separated value plus that range has estimated using confidence interval.

Now, the second assumption that should hold to true is the linear relationship, it should hold true. So, the underlying relationship between the outcome variable and the set of predictors that should hold true, that should be linear and because the first assumption we can recall that it is about that the relationship between outcome variable and the set of predictors is following that what is represented in the regression equation. So, therefore, linear relationship holds true, should hold true otherwise the model would not be predicting values in a reliable fashion.

Observations are independent, so, observations should be independent all the observation that we have, the sample that we have, though these observation should be independent of each other there should not be any dependency and then the last assumption is about the variability in the outcome variable. So, the variability in the outcome variable should be same irrespective of the values of the predictors. So, this particular proper property is also called homoskedasticity. So, this should also hold true that the variability in the outcome variable it should be same.

So, if you want to understand that this particular graph that we have, if we look at the variability of these points the outcome variable which is generally represented on y axis and the predictors which are generally represented on x axis; if we look at the variability of these points that remains same you know irrespective of the values that are being taken on the predictor axis that is on x axis irrespective of the values the if we look at the variability that looks quite similar. So, this is following that last assumption homoskedasticity and therefore, now we could have if this is held true then we can have the unwise predictions and the smallest average squared error.

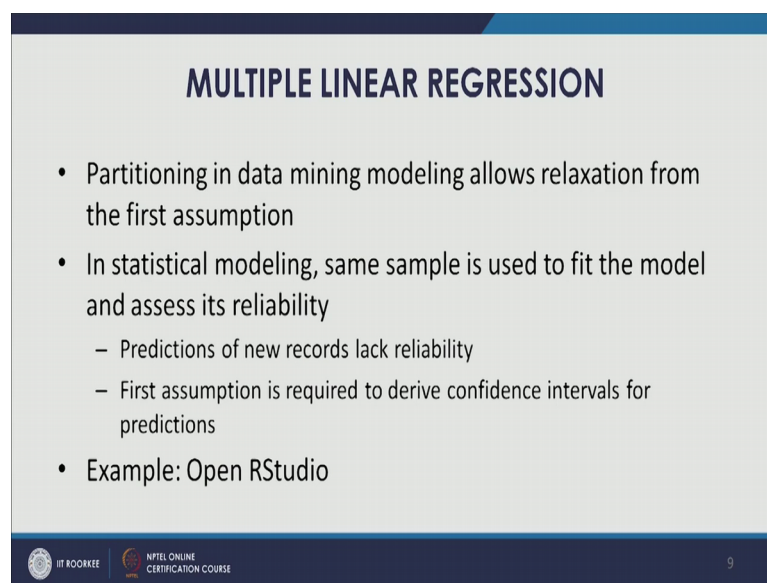
Now, if we look at these assumptions, let us again go back to the assumption the very first assumption that we talked about the noise follows the normal distribution. This is as we discussed this is mainly for statistical modeling that because we use the same sample, but when we talk about data mining approach the partition that we do in data mining modeling that allows relaxation from the first assumption.

So, when we talk about data mining of modeling, let us say this particular bar is representing our sample and generally, we partition the sample into 3 sets. So, because of this partition thing because we are building our model on this particular partition and the model is assessed on the remaining partitions, either validation partition if it is not part of the modeling process or the test partitions. So, because of this we do not the estimates that we get the performance of the model is actually evaluated on performance evaluation happens on this particular partition.

So, therefore, if the model is giving close enough error values the matrix that we use the performance matrix, the numbers that we get from performance matrix if they are quite similar, quite close in both training and validation then probably the model is good and therefore, because we have used different partitions we do not need to follow the first

assumption that is noise follows the normal distribution, that is, mainly for a statistical setting where we have just one sample and the same sample is used for the model building exercise the same sample is used for the evaluation exercise. So, there because the estimates might not be reliable therefore, we need to derive confidence interval. So, so as we are sure about that our estimates are following within that range. So, as the same thing is expressed in the second point of this particular slide that in statistical modeling same sample is used to fit the model and assess to reliability therefore, predictions of new records might lack reliability.

(Refer Slide Time: 26:49)



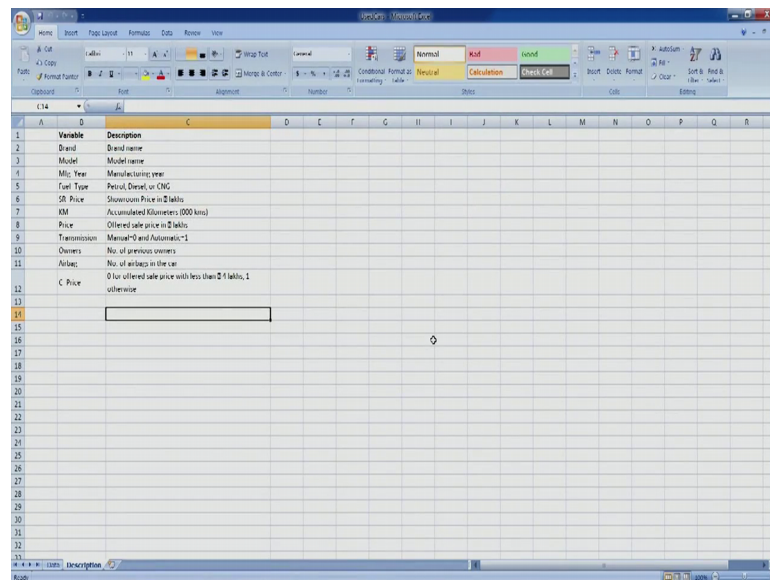
MULTIPLE LINEAR REGRESSION

- Partitioning in data mining modeling allows relaxation from the first assumption
- In statistical modeling, same sample is used to fit the model and assess its reliability
 - Predictions of new records lack reliability
 - First assumption is required to derive confidence intervals for predictions
- Example: Open RStudio

IT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 9

And, therefore first assumption that is required to derive confidence interval for prediction. Now, let us go through an exercise to understand this particular technique.

(Refer Slide Time: 27:07)

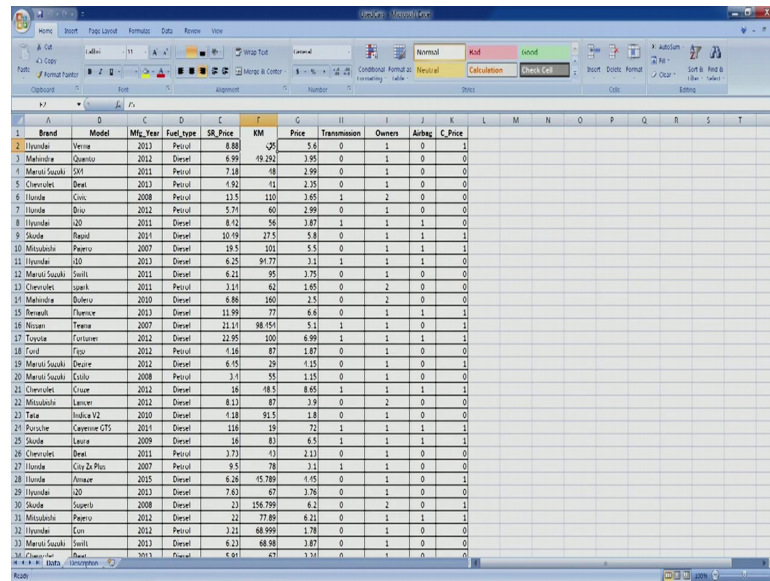


The screenshot shows an Excel spreadsheet with a table of variables and their descriptions. The table has two columns: 'Variable' and 'Description'. The variables listed are: Brand, Model, Mfg. Year, Fuel Type, SR Price, KM, Price, Transmission, Owners, Airbag, and C Price. The descriptions provide details for each variable, such as 'Brand name', 'Model name', 'Manufacturing year', 'Petrol, Diesel, or CNG', 'Showroom Price in ₹ lakhs', 'Accumulated Kilometers (000 kms)', 'Offered sale price in ₹ lakhs', 'Manual=0 and Automatic=1', 'No. of previous owners', 'No. of airbags in the car', and '0 for offered sale price with less than ₹ 1 lakhs, 1 otherwise'.

Variable	Description
Brand	Brand name
Model	Model name
Mfg. Year	Manufacturing year
Fuel Type	Petrol, Diesel, or CNG
SR Price	Showroom Price in ₹ lakhs
KM	Accumulated Kilometers (000 kms)
Price	Offered sale price in ₹ lakhs
Transmission	Manual=0 and Automatic=1
Owners	No. of previous owners
Airbag	No. of airbags in the car
C Price	0 for offered sale price with less than ₹ 1 lakhs, 1 otherwise

So, the data set that we are going to use for this exercise is this used car data set that we have used before. In this we have these variables brand. This data is about used car. So, this is about, the task that we are going to handle is the predict prediction of prices for used cars and based on this historical information about those used cars. The information that we have on these used cars as is brand name, the model name, the manufacturing year, fuel type; it could be petrol diesel or CNG. Then, we have a SR price, that is showroom price in lakhs of rupees and then we have km, that is, accumulated kilometres and thousands of kilometres and then price we have offered sale price in lakhs of rupees and then the whether the car is manual or automatic that is represented by either 0 or 1, then the owners number of previous owners and the airbag, number of airbags in the car and then we have another variable see price that is mainly for the classification task. So, we would not be using this particular variable in our exercise because we would be applying regression modelling which is generally for prediction tasks.

(Refer Slide Time: 28:39)



	A	D	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	Brand	Model	Mfg_Year	Fuel_Type	SR_Price	KM	Price	Transmission	Owners	Airbag	C_Price									
1	Honda	Civic	2013	Petrol	8.88	90	5.6	0	1	0	1									
2	Mahindra	Quantum	2012	Diesel	9.99	49,261	1,95	0	1	0	0									
3	Maruti Suzuki	Swift	2011	Petrol	7.18	48	2.99	0	1	0	0									
4	Chevrolet	Beat	2013	Petrol	1.92	41	2.35	0	1	0	0									
5	Honda	Civic	2008	Petrol	13.5	110	3.65	1	2	0	0									
6	Honda	City	2012	Petrol	5.71	60	2.99	0	1	0	0									
7	Honda	City	2011	Diesel	8.92	54	1.87	1	1	1	0									
8	Skoda	Rapid	2011	Diesel	10.19	27.5	5.8	0	1	1	1									
9	Mitsubishi	Pajero	2007	Diesel	18.5	101	5.5	0	1	1	1									
10	Honda	City	2013	Diesel	6.25	91.77	3.1	1	1	1	0									
11	Maruti Suzuki	Swift	2011	Diesel	6.21	95	3.75	0	1	0	0									
12	Chevrolet	Ignite	2011	Petrol	3.21	62	1.65	0	2	0	0									
13	Mahindra	Quantum	2010	Diesel	6.86	160	2.5	0	2	0	0									
14	Renault	Fluence	2013	Diesel	11.99	77	6.6	0	1	1	1									
15	Nissan	Teana	2007	Diesel	21.11	88.451	5.1	1	1	0	1									
16	Toyota	Fortuner	2012	Diesel	22.95	100	6.99	1	1	1	1									
17	Ford	Figo	2012	Petrol	1.96	87	1.87	0	1	0	0									
18	Maruti Suzuki	Ignite	2012	Diesel	6.45	29	1.15	0	1	0	1									
19	Maruti Suzuki	Ignite	2008	Petrol	3.1	55	1.15	0	1	0	0									
20	Chevrolet	Cruze	2012	Diesel	18	18.5	8.65	1	1	1	1									
21	Mitsubishi	Lancer	2012	Diesel	8.13	67	3.9	0	2	0	0									
22	Toyota	Indica V2	2009	Diesel	1.18	91.5	1.8	0	1	0	0									
23	Porsche	Cayenne GTS	2011	Diesel	118	39	72	1	1	1	1									
24	Skoda	Leora	2009	Diesel	16	83	6.5	1	1	1	1									
25	Chevrolet	Beat	2011	Petrol	3.73	43	2.13	0	1	0	0									
26	Honda	City & Prio	2007	Petrol	9.5	78	3.1	1	1	0	0									
27	Honda	Amaze	2015	Diesel	6.26	45,789	1.45	0	1	0	1									
28	Honda	City	2013	Diesel	7.63	67	3.76	0	1	0	0									
29	Skoda	Superb	2008	Diesel	23	156,799	6.2	0	2	0	1									
30	Mitsubishi	Pajero	2012	Diesel	22	77.89	6.21	0	1	1	1									
31	Honda	City	2012	Petrol	2.21	68,999	1.78	0	1	0	0									
32	Maruti Suzuki	Swift	2013	Diesel	6.23	88.88	1.87	0	1	0	0									
33	Chevrolet	Ignite	2013	Petrol	4.81	67	1.31	0	1	0	0									

So, let us have a look at the data that we have. So, this is the data that we have, as you can see. We have these variables and around 79 observations. Now, with the help of these predictors we want to estimate the price of used cars, this is which is there in the this particular column price. So, all these variables are going to be used in the modeling exercise and the price in this case is the already it is the continuous variable or quantitative variable. So, therefore, we can go ahead with our modeling exercise, there we can apply regression model. So, we will stop here at this point and we will continue our discussions from here will our model in the next class, in the next lecture.

Thank you.